

Benchmarking machine learning algorithm for stunting risk prediction in Indonesia

Nadya Novalina, Ibrahim Amyas Aksar Tarigan, Fatimah Kayla Kameela, Mia Rizkinia
Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia

Article Info

Article history:

Received Jul 13, 2024
Revised Dec 16, 2024
Accepted Dec 25, 2024

Keywords:

Cross industry standard process
for data mining
Machine learning
Stunting
Synthetic minority
oversampling technique
Undersampling

ABSTRACT

Stunting is a condition caused by poor nutrition that results in below-average height development, potentially leading to long-term effects such as intellectual disability, low learning abilities, and an increased risk of developing chronic diseases. One effort to reduce stunting is to apply a machine learning algorithm with a data science approach to develop risk prediction models based on factors in stunting. The study used the current cross industry standard process for data mining (CRISP-DM) framework to gain insight and analyzed 1561 records of data collected from the Indonesia family life survey (IFLS) for the prediction models. Two sampling methods, random undersampling, and oversampling synthetic minority oversampling technique (SMOTE), were employed and compared to overcome the data imbalance problem. Four machine learning classifier algorithms were trained and tested to determine the best-performing model. The experiment results showed that the algorithms yielded an average accuracy of more than 75%. Using the undersampling technique, the accuracy obtained by logistic regression, k-nearest neighbor (KNN), support vector classifier (SVC), and decision tree classifier were 95.21%, 78.91%, 92.97%, and 86.26% respectively. Meanwhile, the oversampling technique reached 96.17%, 88.50%, 93.29%, and 95.21%, respectively. Logistic regression emerges as the best classification, with oversampling yielding superior performance.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Mia Rizkinia
Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia
Kampus UI Depok, Kukusan, Beji, Depok, West Java 16424, Indonesia
Email: mia@ui.ac.id

1. INTRODUCTION

Stunting is a critical issue faced by many developing countries, including Indonesia, and it has significant long-term effects on the growth and development of children. According to the World Health Organization (WHO), stunting is defined as the condition where a child's height falls below two standard deviations (-2 SD) from the median standard of healthy child growth [1]. Its impact extends beyond the physical aspects, affecting children's cognitive development. It also reduces optimal educational performance, potentially diminishes intellectual and motor capacities, and influences economic and social aspects [2]-[4].

The WHO data from 2018-2020 reveals that Indonesia ranks second in the Southeast Asia region for the highest prevalence of stunting, with Timor-Leste in first place and the Philippines in third [3]. According to the Indonesian Health Survey (SKI), conducted in 2023, 21.5% of children under five years old in Indonesia suffer from stunting [5]. This indicates that 1 in 5 children in Indonesia is affected. However, this percentage still exceeds the WHO threshold, which recommends that the prevalence of stunting should be less than 20% [6].

The early stages of a child's development, particularly the first thousand days after birth, constitute a challenging period for growth and development [7], [8]. Stunting, a highly complex problem with contributing factors such as inadequate maternal nutrition, health conditions, suboptimal child-feeding practices, and recurring subclinical infections [1], is often linked to poverty. In areas with high poverty rates, parents frequently face challenges in meeting basic household needs, including providing adequate nutritional intake for their children. Furthermore, the level of education plays a pivotal role in exacerbating nutritional problems. A lack of understanding regarding the importance of proper nutritional practices often leads to inappropriate feeding habits [9]. For example, many parents in Indonesia may poorly grasp the significance of early breastfeeding initiation as a crucial step in ensuring optimal nutrition intake [10].

These findings trigger the need for a deeper understanding of the stunting issue and efforts to address it are crucial to ensure the well-being of the young generation in Indonesia and similar countries. In this context, this research aims to apply a data science and machine learning approach to delve deeper into the factors influencing stunting. This approach allows for meticulous data analysis that identifies key contributors and develops more effective solutions. Accordingly, this research builds upon similar studies that have explored the application of data analysis and machine learning to predict the risk of stunting, providing valuable insights into this complex issue [11], [12].

Chilyabanyama *et al.* [11] compared machine learning algorithms to predict stunting in Zambia, considering factors such as the child's gender, mother's age, household head's age, child's age, household characteristics, maternal employment, education, and family size. They reported that the random forest classifier achieved 79% accuracy. In another study, Shen *et al.* [12] achieved 72.8% accuracy using the XGBoost classifier to predict stunting in Papua New Guinea, focusing on factors including residence in the Highlands Region, child's age, wealth status, and birth size. Both papers collectively revealed limitations in optimally identifying the relationships between essential features, leading to an incomplete understanding of the factors crucial for predicting stunting. Furthermore, existing research has yet to address parameters such as birth weight, prematurity status, breastfeeding, food and protein intake, and consumption expenditure.

In light of this, our research aims to analyze the dominant factors contributing to stunting incidents in Indonesia. Specifically, to address the gaps above, we will utilize the cross-industry standard process for data mining (CRISP-DM) framework during the data analysis stages to manage complex parameters. Advanced analytical techniques and evaluations, along with the implementation of machine learning, will be carried out to create a simple stunting symptom detection model using data from the Indonesia family life survey (IFLS) for the years 2014-2015. The data will be reprocessed based on the tested variables. The processed output data will then be used to build a machine learning model by training the classifier algorithm through supervised learning to assess its performance.

2. METHOD

2.1. Data source

The data used in this study is in the form of longitudinal secondary data collected by the RAND Labor and Population division of the IFLS-5 from 2014 to 2015. This data is openly accessible for general use on the RAND Corporation study website (<https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/ifls5.html>). IFLS data comprises socioeconomic data, household health surveys, community or group surveys, and service provider information in Indonesia. The research was conducted randomly among households residing in 13 out of the 34 provinces in Indonesia. Data was collected continuously for the same households and individuals over time. Since its inception, IFLS has produced six different data sets, with IFLS-5 being the most recent.

The variables used in this study encompass various aspects relevant to understanding the level of stunting in children. Most of these variables have also been used in previous research [9], [13], [14] that categorizes factors into child-related, parental, environmental, and nutritional aspects. In their study, Weingarten *et al* [9] identified height, paternal education, and protein intake expenditure as dominant factors in stunting. Taguri *et al* [13] specifically highlighted the relationship between protein intake, birth weight, parental education, father's occupation, and family economic status with stunting in children. Similarly, Paudel *et al.* [14] found associations between stunting and variables such as mothers without earnings, food deficit households, caretakers other than the mother, kitchens without ventilation, children exposed to pesticides, and breastfeeding. Additionally, they identified other contributing factors, including dietary diversity below WHO standards and diarrhea. These studies collectively emphasize the critical role of these variables in understanding stunting levels among children, offering valuable insights into the multifaceted factors influencing this condition.

2.2. Research workflow

This study employs a data science approach with the CRISP-DM [15] framework, which has proven effective in addressing complex data analysis challenges. CRISP-DM consists of six sequential phases that complement each other, as shown in Figure 1. This approach allows the development of an effective model to formulate recommendations and in-depth inferences about child stunting. The goal is to provide valuable contributions to understanding and handling stunting. The research process involves a series of steps, starting from business understanding and data comprehension through data visualization and statistical analysis, such as p-value, to validate variables according to the research hypotheses. The next stage is data preparation, encompassing data preprocessing, utilizing feature selection to reduce non-significant features, and addressing data imbalances using undersampling and oversampling techniques. The modeling process involves several machine learning models, such as logistic regression, k-nearest neighbors (KNN), support vector classifier (SVC), and decision tree classifier. These models are trained, tested, and evaluated, including comparing the employed sampling techniques. The final step involves testing the model with a confidence score and deploying the model through a graphical user interface (GUI) for result predictions.

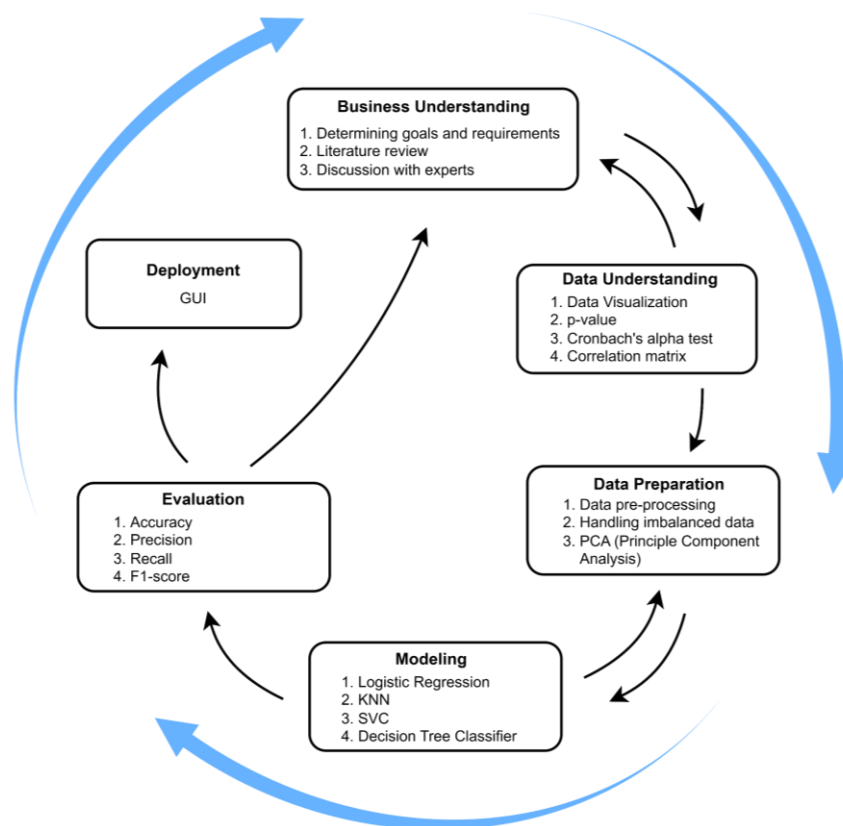


Figure 1. Overview of the proposed model architectures

2.3. Machine learning model

In this stage, the development of a predictive model is carried out using diverse machine learning algorithms. Machine learning is the scientific study of algorithms and statistics that enables computers to perform tasks without explicit instructions by leveraging patterns and inferences [16]. In the context of our study, we tested some machine learning classification algorithms, encompassing:

- Logistic regression: a classification algorithm that examines the relationship between input features and the probability of output results. It is used to classify entities into specific classes, especially when outliers are present in the data. Logistic regression employs maximum likelihood estimation to determine parameters describing the relationship between input features and output probability [17].
- KNN: a non-parametric classification algorithm that relies on the proximity of individual data points to perform classification or prediction. It determines the label of a data point based on the majority vote from its nearest neighbors [18].

- c. SVM: an algorithm that classifies data by creating a linear separator that maximizes the distance between classes. SVM seeks the best hyperplane to separate data into the corresponding classes [19].
- d. Decision tree: this algorithm creates a tree-like structure with internal nodes that partition data into a series of rules, leading to leaf nodes representing the target labels [20].

We selected these algorithms based on the specific needs of our classification task. With a relatively small dataset, our focus is on binary classification for predicting stunting status. Logistic regression was selected for its simplicity and efficiency in handling binary tasks with limited data. KNN adopts a proximity-based approach, which is helpful when dealing with complex relationships between features and stunting risk. SVM is used for its effectiveness in handling both linear and non-linear separations, providing flexibility for capturing intricate patterns. Decision trees are employed for their interpretability, offering clear, rule-based structures that aid in understanding factors contributing to stunting risk. Each algorithm contributes unique strengths to our predictive model, enhancing our assessment of stunting likelihood in children based on the dataset. Proper hyperparameter tuning is essential for model quality [21].

2.4. Model performance evaluation

Confusion matrix was used for the evaluation metric, providing an overview of how well the model's predictions align with the actual outcomes. It consists of four main components: true positive (TP), representing positive data that was correctly predicted; true negative (TN), indicating negative data that was correctly predicted; false positive (FP), encompassing negative data incorrectly predicted as positive; and false negative (FN), covering positive data incorrectly predicted as negative [22]. Based on this, four evaluation metrics are used to measure the classifier's performance. The mathematical [23] describing these metrics are presented in (1)-(4).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1 - score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

3. RESULTS AND DISCUSSION

3.1. Data understanding

A detailed understanding of the variables pivotal to the analysis is provided. Variables are categorized as tested, supported by empirical evidence based on prior research findings [13], [14], or suspected, requiring further investigation. Tested variables include birth weight, premature status, breastfeeding, age, weight, height, parental education and occupation, water source, toilet facilities, and expenditure patterns, as detailed in Table 1. Suspected variables, such as household smoking habits, lack empirical validation but may influence outcomes. By comprehensively understanding these variables, hypotheses can be formulated, and actionable insights can be derived to address the analytical objectives.

Table 1. Potential variables for stunting occurrence

Variable	Description	Tested/suspected variables
Birth weight (kg)	Birth weight of the child	Tested [13], [14]
Premature status	Is the child born premature or normal? 0=Premature, 1=Normal	Tested [13]
Breastfeeding	Did the child receive exclusive breastfeeding? 0=No, 1=Yes	Tested [13]
Age	Age of the child (0-5 years)	Tested [13]
Weight (kg)	Current weight of the child	Tested [13]
Height (cm)	Current height of the child	Tested [13]
Father's education	Years of education completed by the father	Tested [13]
Mother's education	Years of education completed by the mother	Tested [13]
Father's occupation	Does the father work? 0=No, 1=Yes	Tested [13]
Mother's occupation	Does the mother work? 0=No, 1=Yes	Tested [13]
Water	Main source of drinking water? 0=Not protected, 1=Protected	Tested [13]
Toilet	Using a toilet with a septic tank? 0=No, 1=Yes	Tested [13]
Smoking	Is there any household member who smokes? 0=No, 1=Yes	Suspected
Household food	Total expenditure for household food.	Tested [13]
Protein intake	Total expenditure for protein intake	Tested [13]
Price consumption expenditure (PCE)	Per capita expenditure	Tested [13]

3.1.1. Data labeling

In the initial stage, the z-score is calculated to obtain the stunting status label variable. The z-score measures data deviation from its mean, measured in SD units. If the z-score is < -2 , the data is labeled as 1, meaning stunting. Conversely, if the z-score is > -2 , the data is labeled as 0, indicating not stunting. This process helps categorize stunting status based on the SD from the data distribution. The preliminary examination revealed no missing values in the dataset, allowing the analysis to proceed.

3.1.2. Exploratory data analysis

The distribution of the target data has been analyzed to gain insights into its spread. Among the dataset, 266 instances are labeled 1 (stunting), accounting for 17.04% of the total, while 1295 instances are labeled 0 (not stunting), representing 82.96% of the dataset. We also conducted data visualizations to explore how various features relate to the target variable (stunting status). In Figure 2, we present the distribution of numerical features concerning stunting. The plots reveal that stunting is more prevalent among infants with birth weights under 3 kg and those aged 1 year.

Additionally, individuals with below-average weight and height exhibit a higher likelihood of stunting. Parental education levels below 12 years are associated with a dominant presence of stunting. Lastly, lower household food expenditure, protein intake, and PCE are correlated with an increased prevalence of stunting.

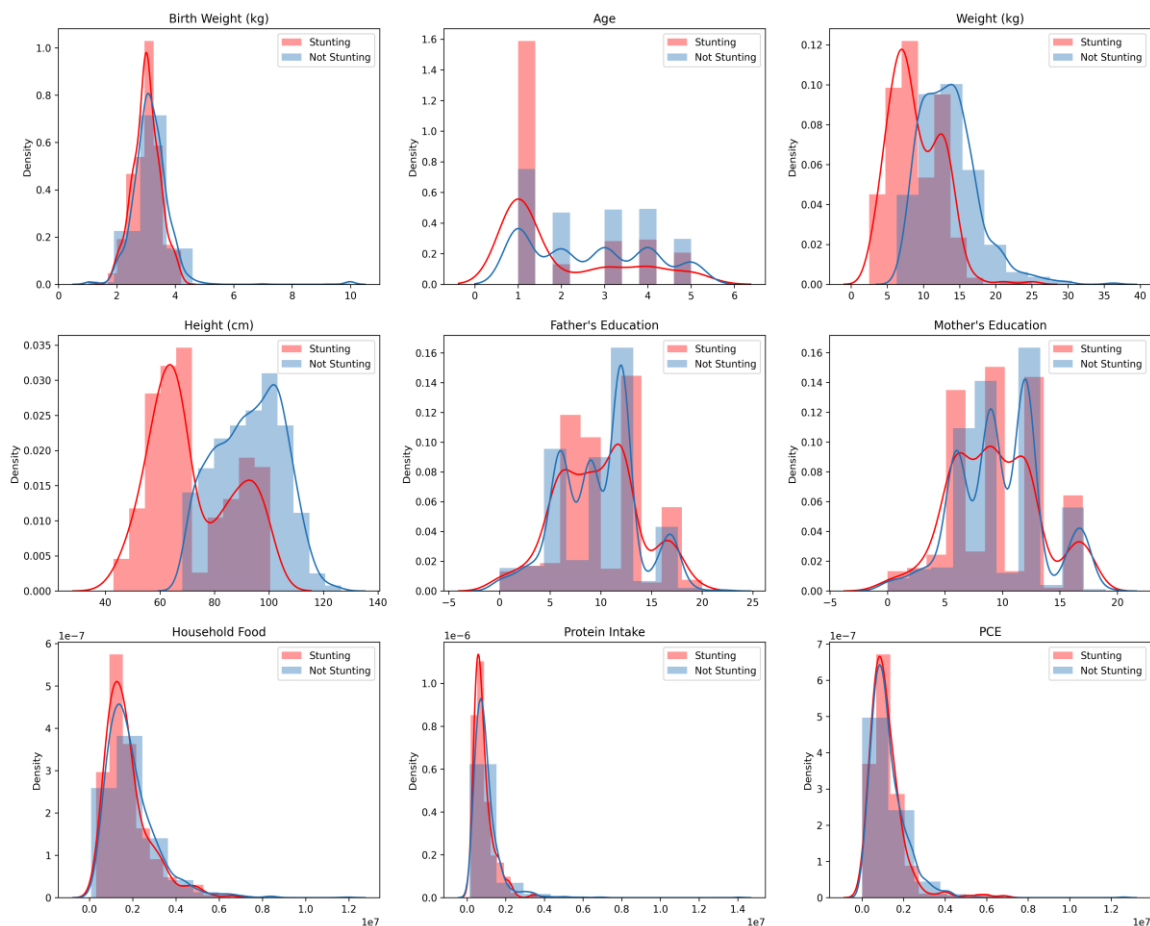


Figure 2. Distribution of numerical features with stunting status

In Figure 3, we present the distribution of categorical features of stunting. The plots reveal that stunting is more common among premature births and exclusive breastfeeding. Additionally, the risk of stunting remains high even when parents are employed. When considering water, toilet, and smoking, stunting remains prevalent despite adequate sanitation. Notably, non-smoking households have lower stunting rates.

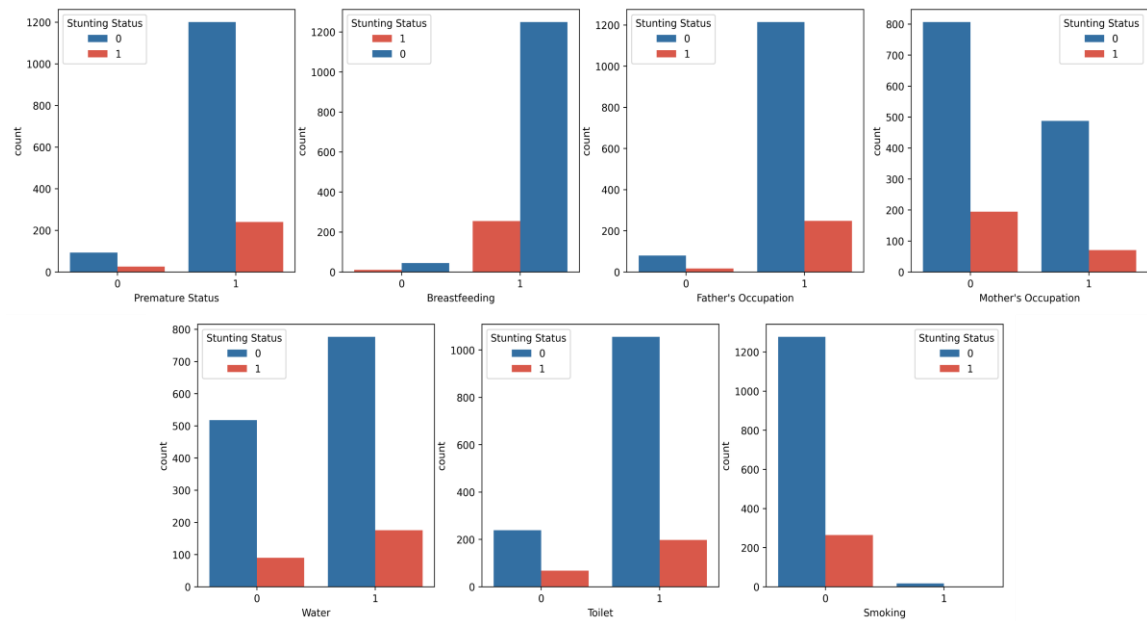


Figure 3. Distribution of categorical features with stunting status

The correlation matrix plays a crucial role in uncovering relationships among variables, as illustrated in Figure 4. This matrix highlights notable correlations between the dependent variables. There is a moderate association between birth weight and premature status (0.45) and stronger correlations, such as age with weight (0.74), age with height (0.84), and weight with height (0.87). Additionally, the father's education shows a moderate positive correlation with the mother's education (0.6). Meanwhile, household food exhibits a strong positive correlation with protein intake (0.67) and PCE (0.75). Notably, protein intake and PCE demonstrate a moderate correlation of 0.44, whereas other dependent variables exhibit weaker correlations.

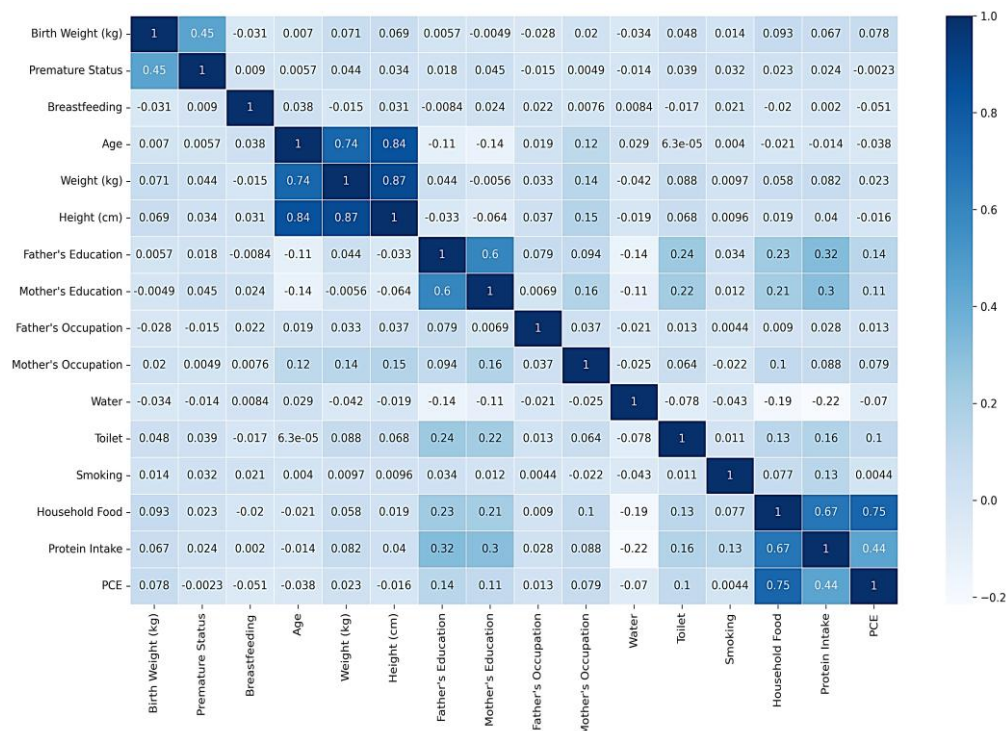


Figure 4. Correlation matrix between variables

3.1.3. P-value analysis

P-value calculations are performed for each variable in the dataset. P-value is a crucial statistical metric in hypothesis testing that aids in understanding the significance of each variable to the research hypothesis. The hypothesized variable is Smoking, derived from the question, "Does any household member smoke?". The initial hypothesis is that if this variable is dominant, the risk of a child experiencing stunting will increase if there is a family member who smokes. Variables with a p-value <0.05 are considered significant, indicating a significant impact on the analysis, while those with a value >0.05 are deemed insignificant [24]. There are five variables with a p-value of <0.05 : age, weight, height, father's education, and mother's education. Notably, the hypothesized variable Smoking is not considered significant as its p-value exceeds 0.05. Three variables, i.e., age, weight, and height, are significant in p-value analysis and strongly correlate with the target data according to the correlation matrix.

3.2. Data preparation

3.2.1. Data preprocessing

The dataset has been examined and found to have no missing values. We split the dataset using a ratio of 80:20. The stratified K-Fold cross-validation (CV) technique, with $K=5$, is applied to the training data to ensure the equitable distribution for robustness and effectiveness of the model.

It is essential to acknowledge that the dataset exhibits class imbalance. Of the data, 266 points are labeled 1 (stunting), representing 17.04% of the dataset, while 1,295 points are labeled 0 (not stunting), constituting 82.96% of the total. To address this class imbalance, we employ two sampling methods on the training data. Random undersampling [25] reduces the number of samples from the majority class, while the synthetic minority over-sampling technique (SMOTE) [26] generates synthetic samples for the minority class. These steps are instrumental in creating a balanced and representative dataset for model training and evaluation.

3.2.2. Feature selection

Our study classifies stunting and aims to predict and evaluate key contributing features. While the original dataset is used for variable explorations, the refined dataset plays a pivotal role in modeling and determining the best model and additional evaluations. This dataset serves as a valuable tool for assessing model performance, evaluating key features contributing to stunting, and guiding evidence-based recommendations.

To address less important variables identified through statistical tests such as p-values and correlation matrices, we employed a method known as feature selection with mutual information from scikit-learn [27]. This method was chosen for its capability to pinpoint the most relevant features for predicting outcomes by assessing the information contributed by each feature in relation to the output class. It is adept at handling both numerical and categorical data. This method is essential for reducing the number of variables in a classification task. Achieved either by discarding less useful ones through feature selection or refining them during data preprocessing [28], [29]. From the initial 16 variables, 10 were selected as the primary factors in stunting occurrence. These variables include height (cm), weight (kg), age, PCE, household food, mother's occupation, protein intake, breastfeeding, birth weight (kg), and premature status.

3.3. Model training and performance evaluation

The training was conducted using logistic regression, KNN, SVC, and decision tree, in which random undersampling and SMOTE were used to handle the imbalance data; then, the results were compared. The hyperparameters were tuned using GridSearchCV. First, we examined the results of random undersampling. Table 2 summarizes that logistic regression consistently outperformed the other models with a CV score of 94.37, accuracy of 95.21, precision of 96.07, recall of 95.21, and an F1-score of 95.41. In summary, random undersampling significantly contributed to addressing class imbalance and improving overall model performance in predicting stunting and non-stunting cases within the context of our study.

Table 2. Model performance metrics using random undersampling

Model	CV score	Accuracy	Precision	Recall	F1-score
Logistic regression	94.37	95.21	96.07	95.21	95.41
KNN	79.35	78.91	86.17	78.91	81.03
SVC	92.27	92.97	94.54	92.97	93.36
Decision tree	84.75	86.26	89.74	86.26	87.28

Next, the results of training with SMOTE were examined. Table 3 demonstrates that logistic regression continued to excel, achieving the highest scores across all evaluation metrics using SMOTE. It achieved a CV score of 94.54, accuracy of 96.17, precision of 96.44, recall of 96.17, and an F1-score of 96.25. This consistent performance reaffirms logistic regression's reliability in making accurate predictions for the stunting dataset, whether under random undersampling or SMOTE. In conclusion, the application of SMOTE played a crucial role in rectifying class imbalance and enhancing overall model performance in predicting stunting and non-stunting cases within the context of our study.

Table 3. Model performance metrics using SMOTE

Model	CV score	Accuracy	Precision	Recall	F1-score
Logistic regression	94.54	96.17	96.44	96.17	96.25
KNN	93.09	88.50	88.33	88.50	88.41
SVC	95.41	93.29	94.10	93.29	93.53
Decision tree	94.69	95.21	95.34	95.21	95.26

While random undersampling and SMOTE share the common goal of addressing class imbalance, they employ distinct strategies. Table 4 presents the key performance metrics for logistic regression (best model) under these two techniques. It shows that logistic regression consistently delivers outstanding results using both techniques, with SMOTE providing slightly higher scores across all evaluation metrics. This reaffirmed the reliability and consistency of logistic regression in making accurate predictions for stunting and non-stunting cases, regardless of the data preprocessing technique employed. The results underscore the potential benefits of SMOTE in terms of improved evaluation scores-particularly F1-score, which is crucial for accurately identifying stunting cases while minimizing FPs.

Table 4. Model performance metrics for logistic regression

Sampling technique	CV score	Accuracy	Precision	Recall	F1-score
Random undersampling	94.37	95.21	96.07	95.21	95.41
SMOTE	94.54	96.17	96.44	96.17	96.25

3.4. Testing sample of potential stunting

The model we developed is more focused on prediction than diagnosis. Since logistic regression converts input features into probability values for binary prediction, the provided probability was used to examine the individual conditions in relation to the stunting prediction. In its application, this model is expected to predict the potential for stunting in an individual so that parents can take appropriate preventive measures. Therefore, we conducted a confidence score test to measure the model's confidence level in predicting input data. The model was tested using data samples with a threshold value between stunting and not stunting status. From this, we aim to demonstrate how important the features are to prevent and promote stunting occurrence. The input data for this testing is listed in Table 5.

Table 5. Sample data for testing potential stunting

Data	Height (cm)	Weight (kg)	Age	PCE	Household food	Mother's occupation	Protein intake	Breast feeding	Birth weight (kg)	Premature status
Data 1	78.5	10.6	2	880500	1561000	0	599000	1	2.78	1
Data 2	80	11	2	1365000	1469000	0	573000	1	2.9	1

Using these datasets, the model initially generated probability predictions. Subsequently, we conducted two different actions for each dataset: for data 1, we significantly increased the PCE, household food, and protein intake variables by 50%. While for data 2, we significantly decreased these variables by 50%. The predicted results, before and after these changes, are presented in Table 6.

Table 6. Model performance metrics for logistic regression

Stage	Data	Prediction (0=not stunting, 1=stunting)	Confidence score (%)	
			0	1
Initial	Data 1	1	23.20	76.80
	Data 2	0	62.63	37.37
Modified	Data 1	1	33.68	66.32
	Data 2	1	46.60	53.39

From Table 6, we observed that after modifying the PCE, household food, and protein intake values, the confidence score for not stunting increased by 10.48% for data 1 and decreased by approximately 16.03% for data 2. These results show that by changing certain variables related to stunting, we can estimate the likelihood of a child being identified as stunted. Higher values for these variables indicate a lower likelihood of stunting, while lower values indicate a higher likelihood.

In the context of stunting, certain variables such as age, birth weight, height, weight, and premature status are inherent and unmodifiable aspects of a child's condition. To address stunting effectively, attention should be directed toward modifiable factors, including PCE, household food, mother's occupation, protein intake, and breastfeeding. Among these, PCE, household food, and protein intake offer the greatest flexibility for substantial change, reflecting parental efforts to improve a child's nutrition, health, and education. Conversely, variables like parental occupation, while changeable, often require considerable time and resources for modification. Additionally, breastfeeding is time-limited by the child's age, ceasing after two years. This aligns with recommendations for exclusive breastfeeding during the first six months, followed by continued breastfeeding alongside complementary foods until age two or older.

3.5. Prediction detection model deployment

After completing the model training and testing processes, we implemented the model as a user-friendly GUI. This GUI was developed using the Python Tkinter library. It allows users to input data based on prior processing with feature selection. Once the data is entered, users can initiate the prediction process by clicking the 'Submit' button. The GUI is presented in Figure 5.

Figure 5. Model deployment in the form of a GUI

4. CONCLUSION

This study examined factors influencing stunting in children aged 0-5 in Indonesia and developed a machine learning model for prediction. The evaluation, utilizing p-value, a correlation matrix, and feature selection, consistently highlighted the importance of age, weight, and height in stunting occurrences. Other factors, such as birth weight, premature status PCE, household food, mother's occupation, protein intake, and breastfeeding, also emerged as significant factors. Notably, the hypothesized variable 'smoking' did not exhibit significance.

The machine learning model achieved over 75% accuracy for all models, with logistic regression outperforming others (96.17% vs. 95.21%) when oversampling was used. This model's ability to predict the risk and likelihood of stunting in children is valuable for risk mitigation and raising awareness. For future research, the variables derived from p-value, correlation matrix, and feature selection evaluations can be employed to create a new questionnaire tailored to these variables. Subsequent studies can then utilize this questionnaire to investigate stunting occurrences.

ACKNOWLEDGMENTS

The authors acknowledge all forms of support from the Universitas Indonesia through the funding of Hibah PUTI Pascasarjana for the years 2022–2023.

FUNDING INFORMATION

This work is supported by Universitas Indonesia under the funding of Hibah PUTI Pascasarjana Tahun 2022-2023 No. NKB-325/UN2.RST/HKP/05.00/2022. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nadya Novalina			✓	✓	✓			✓		✓	✓			
Ibrahim Amyas Aksar		✓	✓		✓				✓		✓			
Tarigan														
Fatimah Kayla Kameela				✓		✓				✓			✓	
Mia Rizkinia	✓	✓		✓		✓	✓			✓		✓		✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

INFORMED CONSENT

This study utilized data obtained from the Indonesia Family Life Survey (IFLS), which conducted the survey and collected information with prior informed consent from the participants. Although the names of the household heads were recorded, no personal identifiers related to the children, who are the subjects of this study, were included. The data were analyzed anonymously, ensuring the confidentiality and privacy of all individuals involved. No additional consent was required for this study.

ETHICAL APPROVAL

The authors state no ethical approval requirements for this study.




DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.




REFERENCES

- [1] World Health Organization, "Global nutrition targets 2025: stunting policy brief," World Health Organization, Geneva, 2014.
- [2] United Nations Children's Fund (UNICEF) and World Health Organization (WHO), "UNICEF-WHO Low birthweight estimates: Levels and trends 2000–2015," World Health Organization, Geneva, 2019.
- [3] United Nations Children's Fund (UNICEF), World Health Organization (WHO), and International Bank for Reconstruction and Development/The World Bank, "Levels and trends in child malnutrition: UNICEF/WHO/The World Bank Group joint child malnutrition estimates: key findings of the 2021 edition," *World Health Organization*, Geneva, 2021.
- [4] S. Grantham-McGregor, Y. B. Cheung, S. Cueto, P. Glewwe, L. Richter, and B. Strupp, "Developmental potential in the first 5 years for children in developing countries," *Series, Child development in developing countries*, vol. 369, no. 9555, pp. 60–70, Jan. 2007, doi: 10.1016/S0140-6736(07)60032-4.
- [5] Ministry of Health of the Republic of Indonesia, "Indonesian Health Survey 2023 in Numbers," *Health Development Policy Agency*, Jakarta, 2024.
- [6] Taufiqurokman, "Equality Strategy for Reducing Stunting Prevalence Rate: Case Study of DKI Jakarta Province," *Jurnal Bina Praja*, vol. 15, no. 3, pp. 495–506, Dec. 2023, doi: 10.21787/jbp.15.2023.495-506.
- [7] M. de Onis and F. Branca, "Childhood stunting: A global perspective," *Maternal & Child Nutrition*, vol. 12, no. S1, pp. 12–26, May 2016, doi: 10.1111/mcn.12231.
- [8] A. J. Prendergast and J. H. Humphrey, "The stunting syndrome in developing countries," *Paediatrics and International Child Health*, vol. 34, no. 4, pp. 250–265, Nov. 2014, doi: 10.1179/2046905514Y.0000000158.
- [9] S. E. Weingarten, K. A. Dearden, B. T. Crookston, M. E. Penny, J. R. Behrman, and D. L. Humphries, "Are household expenditures on food groups associated with children's Future Heights in Ethiopia, India, Peru, and Vietnam?," *International Journal of Environmental Research and Public Health*, vol. 17, no. 13, p. 4739, Jul. 2020, doi: 10.3390/ijerph17134739.
- [10] V. J. Flaherman, S. Chan, R. Desai, F. H. Agung, H. Hartati, and F. Yelda, "Barriers to exclusive breast-feeding in Indonesian hospitals: a qualitative study of early infant feeding practices," *Public Health Nutrition*, vol. 21, no. 14, pp. 2689–2697, Oct. 2018, doi: 10.1017/S1368980018001453.
- [11] O. N. Chilyabanyama *et al.*, "Performance of Machine Learning Classifiers in Classifying Stunting among Under-Five Children in Zambia," *Children*, vol. 9, no. 7, p. 1082, Jul. 2022, doi: 10.3390/children9071082.
- [12] H. Shen, H. Zhao, and Y. Jiang, "Machine Learning Algorithms for Predicting Stunting among Under-Five Children in Papua New Guinea," *Children*, vol. 10, no. 10, p. 1638, Sep. 2023, doi: 10.3390/children10101638.
- [13] A. El Taguri *et al.*, "Risk factors for stunting among under-fives in Libya," *Public Health Nutrition*, vol. 12, no. 8, pp. 1141–1149, Aug. 2009, doi: 10.1017/s1368980008003716.
- [14] R. Paudel, B. Pradhan, R. Wagle, D. Pahari, and S. Onta, "Risk Factors for Stunting Among Children: A Community Based Case Control Study in Nepal," *Kathmandu University Medical Journal*, vol. 10, no. 3, pp. 18–24, Apr. 2013, doi: 10.3126/kumj.v10i3.8012.
- [15] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, The MIT Press, 2015.
- [16] M. Taddy, *Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions*, Ed., 1st ed. New York: McGraw-Hill Education, 2019.
- [17] J. Tolles and W. J. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," *JAMA Guide to Statistics and Methods*, vol. 316, no. 5, p. 533, Aug. 2016, doi: 10.1001/jama.2016.7653.
- [18] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, 2019, pp. 1255–1260, doi: 10.1109/ICCS45141.2019.9065747.
- [19] H. Wang, J. Xiong, Z. Yao, M. Lin, and J. Ren, "Research Survey on Support Vector Machine," in *Proceedings of the 10th EAI International Conference on Mobile Multimedia Communications*, EAI, 2017, doi: 10.4108/eai.13-7-2017.2270596.
- [20] F. -J. Yang, "An Extended Idea about Decision Trees," *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 2019, pp. 349–354, doi: 10.1109/CSCI49370.2019.00068.
- [21] M. Claesen and B. D. Moor, "Hyperparameter Search in Machine Learning," *Machine Learning*, Feb. 2015, doi: 10.48550/arXiv.1502.02127.
- [22] J. D. Novaković, A. Veljović, S. S. Ilić, Ž. Papić, and M. Tomovic, "Evaluation of Classification Models in Machine Learning," *Theory and Applications of Mathematics & Computer Science*, vol. 7, no. 1, pp. 39–, 2017.
- [23] H. Dalianis, "Evaluation Metrics and Evaluation," in *Clinical Text Mining*, pp. 45–53, 2018, doi: 10.1007/978-3-319-78503-5_6.
- [24] R. L. Wasserstein and N. A. Lazar, "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, Apr. 2016, doi: 10.1080/00031305.2016.1154108.
- [25] J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data," *2015 IEEE International Conference on Information Reuse and Integration*, San Francisco, CA, USA, 2015, pp. 197–202, doi: 10.1109/IRI.2015.39.
- [26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
- [27] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011.
- [28] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," in *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994, doi: 10.1109/72.298224.
- [29] B. C. Ross, "Mutual Information between Discrete and Continuous Data Sets," *PLoS One*, vol. 9, no. 2, p. e87357, Feb. 2014, doi: 10.1371/journal.pone.0087357.





BIOGRAPHIES OF AUTHORS

Nadya Novalina    received her bachelor's degree in biomedical engineering from Universitas Indonesia in 2022. Her research interests include artificial intelligence, generative AI, and data science, with a particular emphasis on their applications in healthcare. She is dedicated to expanding her expertise in these areas through ongoing research and professional development. She can be contacted at email: nadya.novalina@ui.ac.id.







Ibrahim Amyas Aksar Tarigan    received a Master of Engineering degree from Universitas Indonesia in 2022, with a specialization in Data engineering and business intelligence. Research focuses on data science, machine learning, and data analytics. He can be contacted at email: ibrahim.amyas@ui.ac.id.



Fatimah Kayla Kameela     received her bachelor's degree in biomedical engineering from Universitas Indonesia in 2024. Her research interests include neuroscience, machine learning, and related disciplines, driven by a strong passion for advancing healthcare and its innovations. She can be contacted at email: fatimah.kayla@ui.ac.id.



Mia Rizkinia     received the B.E. and M.E. degrees from Universitas Indonesia, in 2008 and 2011, respectively, and the Ph.D. degree in information engineering from The University of Kitakyushu, Japan, in 2018. Since 2011, she has been with the Faculty of Engineering, Universitas Indonesia, as a Lecturer. She joined the Artificial Intelligence and Data Engineering (AIDE) Research Center, Faculty of Engineering, Universitas Indonesia, as a researcher. Her research interests include image processing, computer vision, remote sensing, and data science. She can be contacted at email: mia@ui.ac.id.